Régression inverse par tranches sur flux de données

Marie Chavent, Stéphane Girard, Vanessa Kuentz, Benoit Liquet, Thi Mong Ngoc Nguyen, Jérôme Saracco

Régression inverse par tranches sur flux de données

Marie Chavent[1,2], Stéphane Girard[3], Vanessa Kuentz[4], Benoit Liquet[5], Thi Mong Ngoc Nguyen[6] & Jérôme Saracco[1,2]

[1] *IMB, UMR CNRS 5251*
Université de Bordeaux / Institut Polytechnique de Bordeaux,
351 cours de la libération, 33405 Talence Cedex, France
e-mail: {marie.chavent, jerome.saracco}@math.u-bordeaux1.fr
[2] *INRIA Bordeaux Sud-Ouest, CQFD team, France*
[3] *INRIA Rhône-Alpes et LJK, France Inovallée, 655, av. de l'Europe, Montbonnot, 38334 Saint-Ismier cedex, France*
e-mail: Stephane.Girard@inria.fr
[4] *Irstea, UR ADBX Aménités et Dynamiques des Espaces Ruraux,*
50 avenue de Verdun Gazinet Cestas, F-33612 France
e-mail: vanessa.kuentz-simonet@irstea.fr
[5] *INSERM U897, ISPED*
Université Victor Segalen Bordeaux 2, 146 rue Leo Saignat, 33076 Bordeaux Cedex
e-mail: Benoit.Liquet@isped.u-bordeaux2.fr
[6] *IRMA UMR 7501*
Université de Strasbourg, 7 rue René-Descartes, 67084 Strasbourg Cedex, France
e-mail: nguyen@math.unistra.fr

Résumé. Dans cette communication, nous nous concentrons sur des données arrivant séquentiellement par bloc. Nous supposons la présence d'un modèle semi-paramétrique sous-jacent incluant une direction EDR (Effective Dimension Reduction) commune β dans chaque bloc. Nous proposons une approche SIR (Sliced Inverse Regression pour régression inverse par tranches) adaptative afin d'estimer β. L'estimateur proposé est plus rapide qu'une application séquentielle de la méthode SIR à l'union des blocs disponibles. Nous montrons la convergence en probabilité et la normalité asymptotique de cet estimateur. Dans une simulation, nous illustrons le bon comportement numérique de notre approche. Nous fournissons également des graphiques permettant de détecter s'il existe une dérive de la direction EDR ou bien des blocs de données aberrantes, et nous illustrons notre approche avec différents scénarios. Quelques extensions possibles de cette méthode sont discutées en conclusion.

Mots-clés. Sliced Inverse Regression (SIR), flux de données.

Abstract. In this communication, we consider block-wise evolving data streams. When a semiparametric regression model involving a common dimension reduction direction β is assumed for each block, we propose an adaptive SIR (Sliced Inverse Regression) estimator of β. This estimator is faster than usual SIR applied to the union of all the blocks, both from computational complexity and running time points of view. We show

the consistency of our estimator at the root-n rate and give its asymptotic distribution. In a simulation study, we illustrate the good numerical behavior of the estimator. We also provide a graphical tool in order to detect if there exists a drift of the dimension reduction direction or some aberrant blocks of data. We illustrate our approach with various scenarios. Finally, possible extensions of this method are given.

Keywords. Sliced Inverse Regression (SIR), data stream.

1 Introduction

Nous considérons le modèle de régression semi-paramétrique à un seul indice proposé par Duan et Li (1991) :

$$Y = f(X'\beta, \epsilon) \tag{1}$$

où la variable à expliquer Y est à valeurs dans $\mathbb{R}$ et la covariable X appartient à $\mathbb{R}^p$ (avec $\mathbb{E}(X) = \mu$ et $\mathbb{V}(X) = \Sigma$). Le terme aléatoire d'erreur ϵ est indépendant de X, de loi arbitraire et inconnue. La fonction de lien f et le vecteur $\beta \in \mathbb{R}^p$ sont inconnus. Le paramètre β n'est pas totalement identifiable dans ce modèle. Ainsi nous nous intéressons à chercher le sous-espace linéaire engendré par β : ici dans le cadre d'un modèle à un seul indice, nous parlons de direction EDR (Effective Dimension Reduction).

Dans cette communication, nous nous focalisons sur des données arrivant séquentiellement par bloc. Nous supposons que chaque bloc t de données est composé des observations $\{(X_i, Y_i), i = 1, \ldots, n_t\}$ qui sont indépendantes et identiquement distribuées (i.i.d). Notre objectif est d'estimer séquentiellement la direction EDR à l'arrivée de chaque nouveau bloc d'observations.

Une approche simple et directe consiste à mettre en commun tous les blocs disponibles et estimer ensuite la direction EDR par la méthode SIR introduite par Li (1991). Un inconvénient de cette méthode est le problème de stockage puisque la taille de l'ensemble de données augmente considérablement avec le nombre de blocs. Afin d'éviter ce problème, nous proposons une méthode SIR adaptative inspirée de l'approche SIR pour population stratifiée introduite par Chavent *et al* (2011). L'idée consiste à combiner convenablement les directions EDR obtenues dans chaque bloc afin d'estimer ensuite la direction EDR commune pour la totalité de blocs disponibles. Avec cette approche, nous n'avons plus besoin de stocker tous les blocs mais seulement leurs directions EDR.

Cette méthode SIR adaptative, appelée SIRds (pour SIR data stream) a été mise en œuvre sur des données réelles pour évaluer les propriétés physiques de la surface de la planète Mars à partir d'images hyper-spectrales, voir Bernard-Michel *et al* (2009) pour une description des données.

2 Estimateur SIR adaptatif

Rappel de la méthode SIR dans le bloc t. La méthode SIR repose sur la condition de linéarité suivante :

$$(C): \quad \forall b \in \mathbb{R}^p, E(X'b|X'\beta) \text{ est linéaire en } X'\beta,$$

qui est vérifiée lorsque X suit une distribution elliptique. Par ailleurs, pour des données de grande dimension (i.e. théoriquement une dimension p de la covariable X tendant vers l'infini), cette condition est presque sûrement vérifiée. Le lecteur peut se référer à Hall et Li (1993) pour plus de détails.

Soit $T(.)$ une transformation monotone de Y. Sous la condition (C) et le modèle semiparamétrique (1), Li (1991) a montré que la courbe de régression inverse centrée $(y \mapsto \mathbb{E}(X|T(Y) = y) - \mu)$ appartient au sous-espace linéaire de dimension 1 de $\mathbb{R}^p$ engendré par $\Sigma\beta$. Une conséquence est que le vecteur propre b_t associé à la valeur propre non nulle de $\Sigma^{-1}\Gamma_t$ est une direction EDR (i.e. est colinéaire à β) où $\Gamma_t = V(E(X|T(Y)))$. Pour obtenir un estimateur de Γ_t qui peut être facilement utilisé en pratique, Li (1991) a proposé un choix de $T(.)$ particulier appelé "tranchage" : il s'agit d'une discrétisation de Y en découpant le support de Y en $H_t \geq 2$ tranches distinctes $s_1, \ldots, s_{H_t}$. En notant $p_h = P(Y \in s_h)$ (resp. $m_h = E(X|Y \in s_h)$) le poids (resp. la moyenne) de la $h^{\text{ème}}$ tranche, la matrice Γ_t peut s'écrire :

$$\Gamma_t = \sum_{h=1}^{H_t} p_h(m_h - \mu)(m_h - \mu)'.$$

Il est alors facile de l'estimer en substituant les moments empiriques aux moments théoriques. Soit $\widehat{\Gamma}_t$ cet estimateur. On obtient donc l'estimateur $\widehat{b}_t$ de la direction EDR en considérant le vecteur propre de $\widehat{\Sigma}^{-1}\widehat{\Gamma}_t$ associé à la plus grande valeur propre où $\widehat{\Sigma}$ est un estimateur de Σ.

Version sur population de l'approche SIRds. Pour chaque bloc t, notons b_t la direction EDR correspondante. Afin de combiner ces différentes directions EDR lorsque T blocs sont disponibles, nous considérons la matrice suivante :

$$M_T = \sum_{t=1}^{T} w_t b_t b_t' \cos^2(b_t, b_T),$$

où les w_t sont des poids positifs tels que $\sum_{t=1}^{T} w_t = 1$. Remarquons que sous les hypothèses de notre modèle, le terme $\cos^2(b_t, b_T)$ est égal à 1 vu que tous les vecteurs b_t sont colinéaires à β. Il est facile de montrer que le vecteur propre principal (i.e. associé à la plus grande valeur propre) v_T de M_T est une direction EDR.

Remarquons pour cela qu'il est possible de reformuler cette approche comme un problème d'optimisation:

$$v_T := \arg\max_{v\in\mathbb{R}^p} \sum_{t=1}^{T} \tilde{w}_t \cos^2(b_t, v) \quad \text{s.c. } ||v|| = 1, \tag{2}$$

avec $\tilde{w}_t = w_t \cos^2(b_t, b_T)$. En effet, en prenant $||b_t|| = 1$, nous avons donc :

$$\sum_{t=1}^{T} \tilde{w}_t \cos^2(b_t, v) = \sum_{t=1}^{T} \tilde{w}_t < b_t, v >^2 = \sum_{t=1}^{T} \tilde{w}_t v' b_t b_t' v = v' \left(\sum_{t=1}^{T} \tilde{w}_t b_t b_t' \right) v = v' M_T v,$$

Ainsi ce problème de maximisation (2) peut se reécrire comme

$$\arg\max_{v\in\mathbb{R}^p} \frac{v' M_T v}{v' v}, \tag{3}$$

la solution de (3) étant clairement le vecteur propre principal de M_T.

Version sur échantillon pour SIRdatastream. Pour $t = 1, \ldots, T$, notons $\widehat{b}_t$ l'estimateur de la direction EDR calculé sur le bloc t . L'estimateur $\widehat{v}_T$ de la direction EDR v_T avec l'approche SIRds est le vecteur propre principal de la matrice $p \times p$ définie par

$$\widehat{M}_T = \sum_{t=1}^{T} w_t \widehat{b}_t \widehat{b}_t' \cos^2(\hat{b}_t, \hat{b}_T) \tag{4}$$

où $w_t = \frac{n_t}{\sum_{j=1}^{T} n_j}$ et $\cos^2(\widehat{b}_t, \widehat{b}_T) = \frac{(\widehat{b}_t' \widehat{b}_T)^2}{(\widehat{b}_t' \widehat{b}_t) \times (\widehat{b}_T' \widehat{b}_T)}$.

Résultats asymptotiques. Les hypothèses suivantes sont nécessaires pour obtenir les résultats asymptotiques pour l'estimateur $\widehat{v}_T$. On considère un nombre de blocs T fixé et une taille de l'échantillon global n qui tend vers l'infini. Soit $n_{h,t}$ le nombre d'observations dans la $h^{\text{ème}}$ tranche dans le bloc t et soit $n_t = \sum_{h=1}^{H_t} n_{h,t}$ le nombre d'observations dans le bloc t.

- (A1) Chaque bloc t est un échantillon indépendant issu du modèle (1).
- (A2) Pour chaque bloc t, le support de Y est partionné en un nombre de tranches H_t fixé telles que $p_h > 0, h = 1, \ldots, H_t$.
- (A3) Pour $t = 1, \ldots, T$ et $h = 1, \ldots, H_t$, $n_{h,t} \to \infty$ (on a donc $n_t \to \infty$ et $n \to \infty$).

Théorème 1 *Pour chaque bloc t et sous les hypothèses (C), (A1)-(A3), nous avons :*

$$\widehat{v}_T = v_T + O_p(n^{-1/2}).$$

La direction EDR estimée $\widehat{v}_T$ converge donc en probabilité vers la direction EDR v_T (colinéaire à β). De plus, nous avons :

$$\sqrt{n}(\hat{v}_T - v_T) \longrightarrow_d W \sim \mathcal{N}(0, \Gamma_W),$$

où l'expression de Γ_W est donnée dans Chavent et al (2012).

3 Quelques résultats de simulation

Pour chaque bloc de données, nous considérons le modèle de régression suivant :

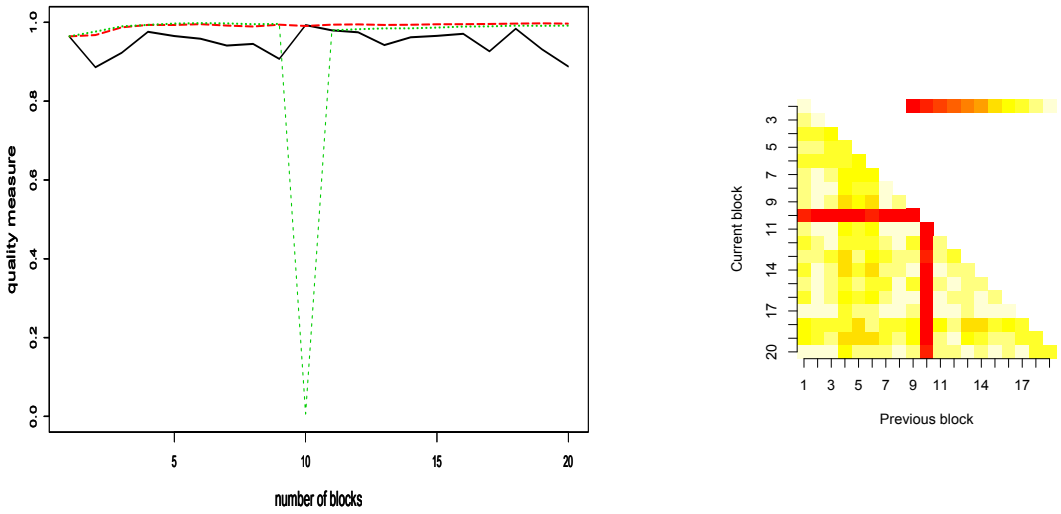
$$Y = (X'\beta_t)^3 + \epsilon, \tag{5}$$

où X suit la loi multinormale $\mathcal{N}_p(0_p, \Sigma)$ avec la covariance Σ choisie arbitrairement (avec $p = 10$), et ϵ suit la distribution normale $\mathcal{N}(0, 1)$ et indépendant de X. Le paramètre β a été indicé par t et peut éventuellement changer au cours du temps. Dans l'exemple ci-après, nous illustrons le caractère adaptatif de l'approche SIRds proposée. Nous montrons le bon comportement de notre approche SIRds par rapport à SIR (sur l'union des blocs disponibles) sur deux scénarios avec des flux de $T = 20$ blocs de tailles $n_t = 200$:

- *Scénario 1* : on considère $T - 1$ blocs avec $\beta_t = (1, -1, 2, -2, 0, \ldots, 0)'$, $t \neq 10$ et le $10^{\text{ème}}$ bloc aberrant avec $\beta_{10} = (1, 1, \ldots, 1)'$.
- *Scénario 2* : on considère $\beta_t = (1, -1, 2, -2, 0, \ldots, 0)'$ pour $t = 1, \ldots, 9$ et puis $\beta_t = (1, 1, \ldots, 1)'$ pour $t = 10, \ldots, 20$.

À chaque instant t (i.e. lorsque les t premiers blocs sont disponibles), nous estimons la direction EDR avec les approches SIRds, SIR basée sur l'union de ces t premiers blocs (appelée SIRu dans la suite) et SIR basée uniquement sur le $t$ème bloc. Cette dernière approche permet non seulement d'illustrer la capacité de SIR à retrouver la direction EDR dans chaque bloc, mais aussi de montrer l'intérêt de combiner l'information en provenance des blocs disponibles afin d'améliorer l'estimation de la direction EDR. Pour chaque scénario, nous avons tracé la mesure $\cos^2(\widehat{\beta}_t, \beta_t)$ de la qualité des estimateurs $\widehat{\beta}_t$ pour chacune des trois approches. Plus la mesure est proche de 1, meilleure est l'estimation. Afin de détecter si une dérive survient ou si des blocs aberrants apparaissent dans le flux de données pour les utilisateurs, nous représentons une image des poids $\cos^2(\widehat{b}_t, \widehat{b}_T)$ utilisés dans le calcul de $\widehat{M_T}$. Plus la couleur est claire (jaune), plus le poids est important (i.e. le nouveau bloc apporte la même information sur la direction de β que les blocs précédents).

Scénario 1 : la direction EDR dans le $10^{\text{ème}}$ bloc diffère de celle des autres blocs



Scénario 2 : changement de direction après le bloc 10

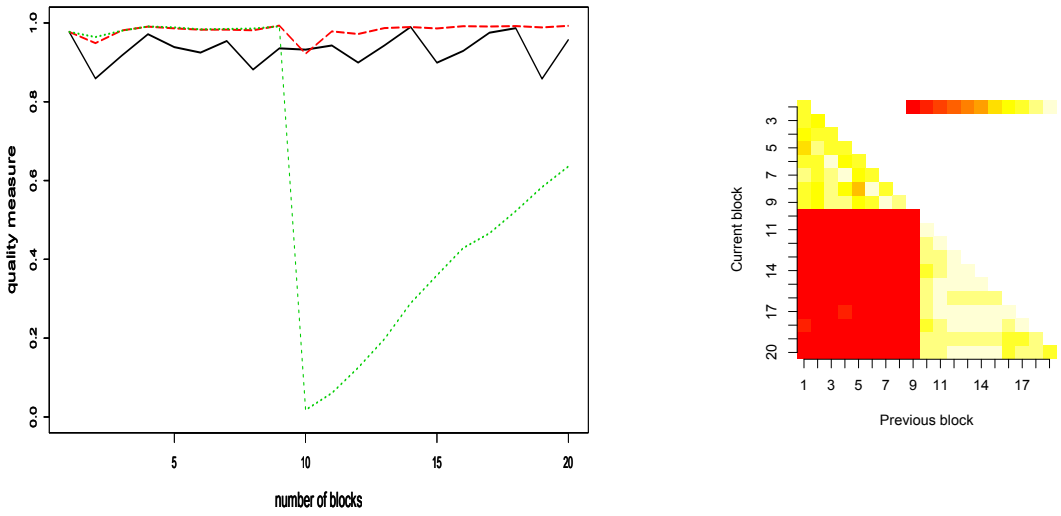


Figure 1: Comportement numérique des estimateurs de SIRds, SIRu et SIR. L'image à gauche représente la qualité $\cos^2(\widehat{\beta}_t, \beta_t)$, en fonction du nombre de blocs (en tirets rouges pour SIRdatastream, en pointillés vert pour SIRu, en noir pour SIR sur uniquement le bloc t). L'image à droite représente les poids $\cos^2(\widehat{b}_t, \widehat{b}_T)$ utilisés dans le calcul de $\widehat{M}_T$.

Plus la couleur est rouge, plus le poids correspondant est faible (i.e. la direction dans ce bloc est différente de celle des autres blocs).

Pour le scénario 1, SIRds et SIRu sont très performantes sur les 9 premiers blocs. Pour le $10^{\text{ème}}$ bloc, SIRu n'est pas capable de retrouver la bonne direction EDR contrairement à SIRds qui s'adapte bien à ce changement de direction. Par ailleurs, l'image du poids indique clairement que ce $10^{\text{ème}}$ bloc diffère des autres.

Pour le scénario 2, l'image du poids montre clairement qu'il y a deux groupes de blocs apportant deux directions différentes (les 9 premiers blocs correspondant au triangle jaune supérieur, et les 11 derniers blocs correspondant au triangle jaune inférieur). L'estimateur de la direction EDR par SIRds reste efficace sur l'ensemble du flux de données contrairement à l'approche SIRu qui ne prend pas en compte le changement de direction après le bloc 10.

4 Conclusion

L'approche SIRds proposée se comporte bien sur des données simulées. Cette approche est bien plus rapide en temps de calcul que l'approche SIR séquentielle. Il est possible d'étendre cette approche à un modèle à plusieurs indices. Il est également possible d'utiliser des méthodes alternatives à SIR comme SIR-II, SAVE ou SIR_α par exemple. La méthode SIRds proposée a été utilisée pour évaluer les propriétés physiques de la surface de la planète Mars à partir d'images hyperspectrales, voir Chavent*et al* (2012).

Bibliographie

[1] Bernard-Michel, C., Douté, S., Fauvel, M., Gardes, L. and Girard, S. (2009), Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression. *Journal of Geophysical Research - Planets* 114, E06005.

[2] Chavent, M., Kuentz, V., Liquet, B. and Saracco, J. (2011), A sliced inverse regression approach for a stratified population. *Communications in statistics - Theory and methods*, 40, 1-22.

[3] Chavent, M., Girard, S., Kuentz, V., Liquet, B., Nguyen, T.M.N. and Saracco, J. (2012), A sliced inverse regression approach for data stream. *Submitted paper*.

[4] Duan, N., Li, K.C. (1991), Slicing regression: a link-free regression method. *The Annals of Statistics*, 19, 505-530;

[5] Hall, P. and Li, K. C. (1993) On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, 21, 867-889.

[6] Li, K.C. (1991) Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86, 316-342.